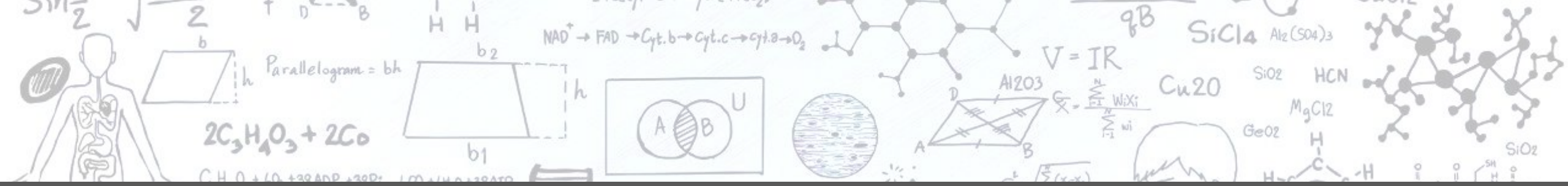


DSW
DATA STEWARDSHIP WIZARD

Research Data Management

Proč a jak?

elixir CZECH REPUBLIC

FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

doc. Ing. Robert Pergl, Ph.D.
robert.pergl@fit.cvut.cz
robert.pergl@ds-wizard.org

Představení

doc. Ing. Robert Pergl, Ph.D.

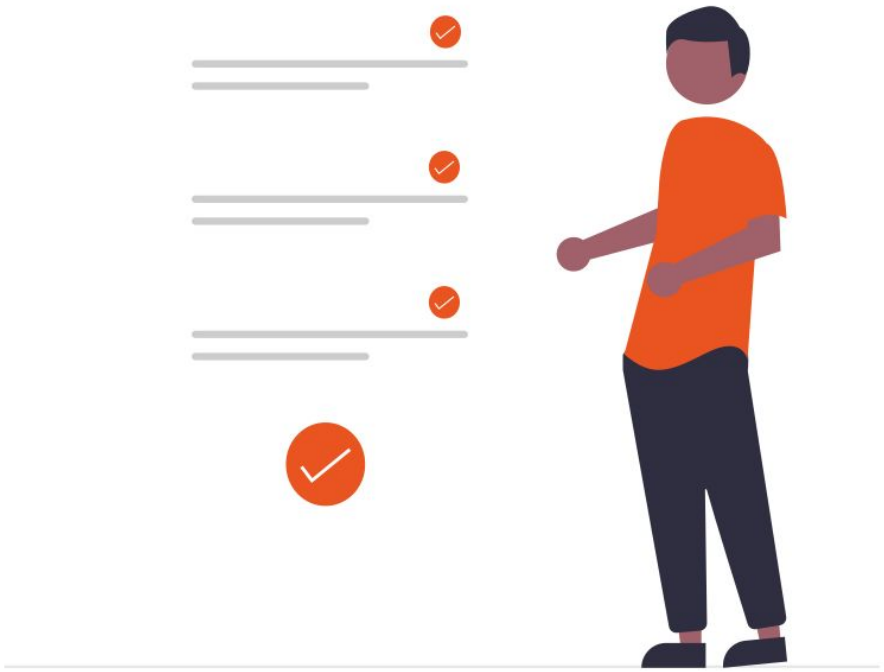
- vedoucí Centra pro konceptuální modelování a implementace na FIT ČVUT (CCMi)
- zástupce ve Výboru ELIXIR CZ za ČVUT
- projektový koordinátor *Data Stewardship Wizard*

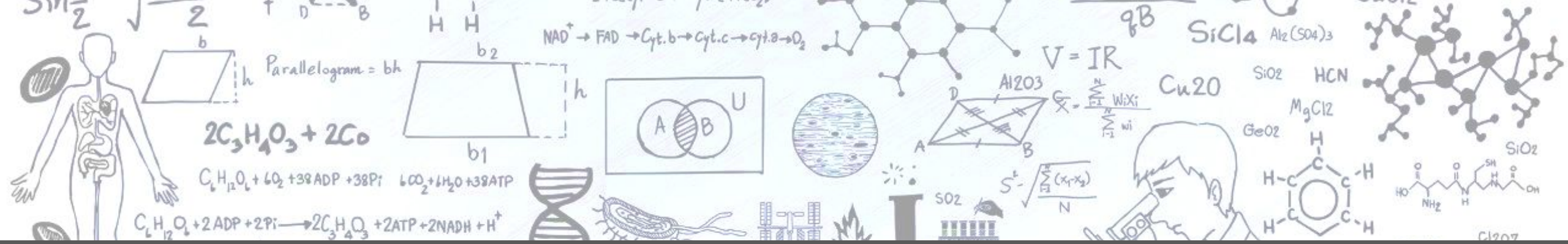
robert.pergl@fit.cvut.cz

robert.pergl@ds-wizard.org

Obsah

- Research Data Management
- FAIR
- Data Management Plan
- Data Stewardship Wizard

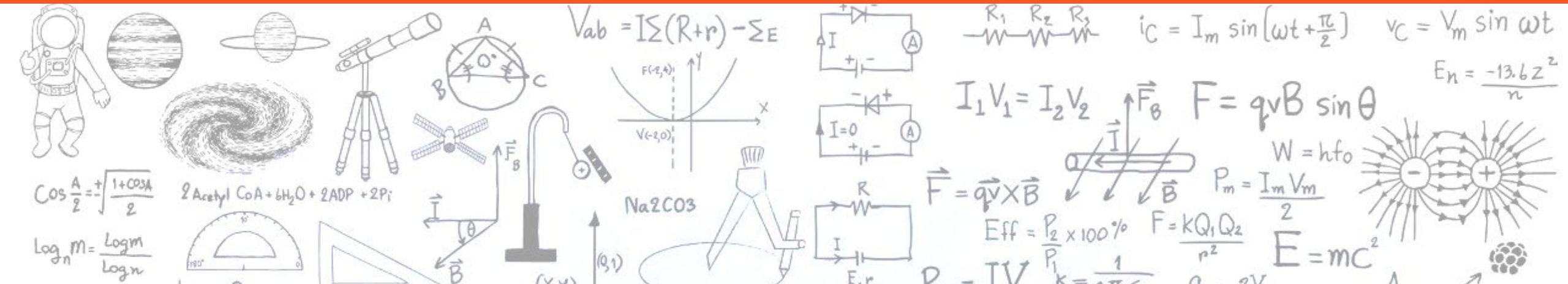




DSW

DATA STEWARDSHIP WIZARD

Research Data Management



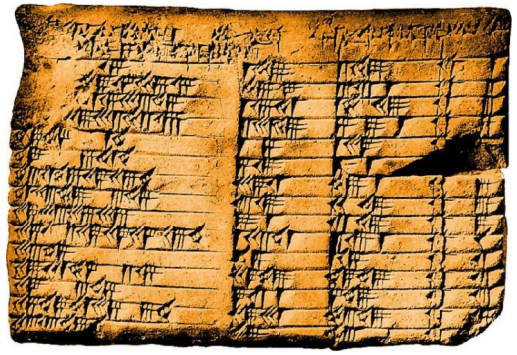
Trocha motivační historie…

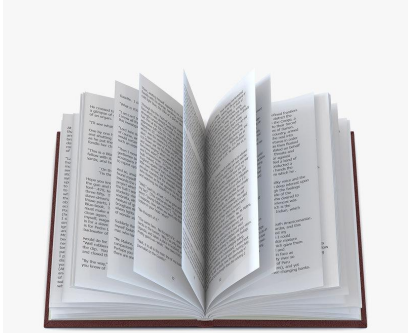
Práce s daty (a zejména jejich sdílení!) stojí za úspěchem Homo Sapiens.

[Why did humans prevail?]

Trocha motivační historie…

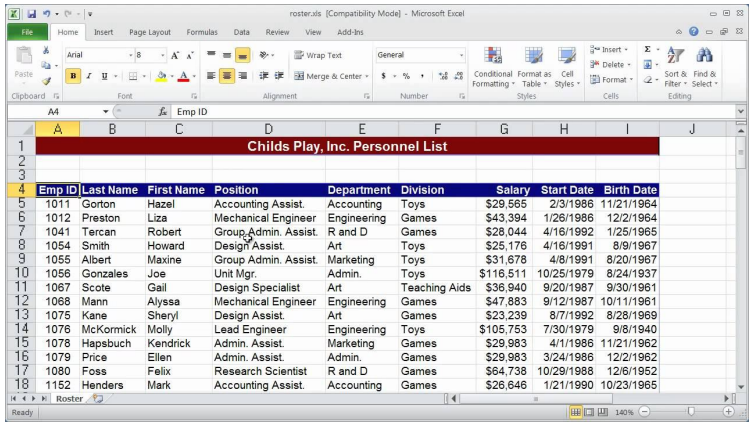
Historie lidstva je i historií sběru, reprezentace, interpretace, zpracování, ukládání, využívání a sdílení dat.

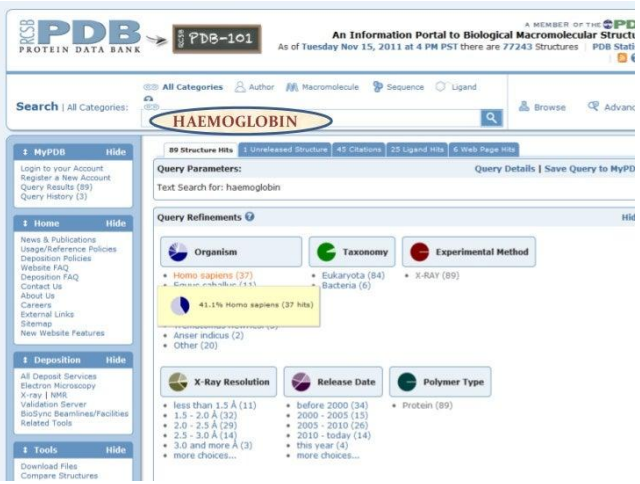




Trocha motivační historie…

Historie lidstva je i historií sběru, reprezentace, interpretace, zpracování, ukládání, využívání a sdílení dat.





Trocha motivační historie…

A též rozvojem systémů správy dat (data management)

Současná situace

- Výrazně rostoucí objem vědeckých dat představuje technické, ale i organizační výzvy.
- Data přináší společnosti výraznou hodnotu – jsou podkladem pro analýzy, rozhodování (viz např. COVID).
- Data se stávají klíčovým motorem pokroku (nejen) ve vědě.
- Správa dat je tak jedním z hlavních aktuálních témat v administrativě vědeckých projektů.

Životní cyklus dat (Data Life Cycle)

https://rdmkit.elixir-europe.org

Data Life Cycle: Plan

- Plánování jakým způsobem se bude během projektu nakládat s daty
- Výstupem by měl být **Data Management Plan** (více dále)

Data Life Cycle: Collect

- Sběr nových dat (metody se liší podle výzkumné oblasti)
- Použití existujících dat (např. z předchozích projektů)
- Důraz na kvalitu použitých dat
- Zaznamenání původu *(provenance)* dat - kdo, pomocí čeho (nástroje), podmínky experimentu, atd.

Data Life Cycle: Process

- Převod dat z uloženého formátu do formátu vhodného pro analýzu
- Vyřazení špatných dat nebo dat z nízkou kvalitou
- Pseudonimizace/anonymizace citlivých dat

Data Life Cycle: Analyse

- Zkoumání nasbíraných dat
- Hlavní část výzkumu – získávání nových znalostí
- Workflow použité v analýze by mělo být reprodukovatelné
- Analýza velkých dat může vyžadovat velký výpočetní výkon
- Potřeba specializovaných oborových softwarových nástrojů (viz např. bio.tools)
- K dispozici jsou velké výzkumné infrastruktury, např. ELIXIR, CLARIN

Data Life Cycle: Preserve

- Zajištění dlouhodobého uchování dat po ukončení projektu:
  - Možnost ověření výsledků projektu i po letech
  - Využití dat v budoucnu pro jiné účely (výuka, další výzkum)

Data Life Cycle: Share

- Sdílení dat s ostatními (např. na dalším výzkumném projektu)
- Sdílení neznamená, že data musí být vždy veřejně dostupná (open), mohou být sdílená pouze za určitých (omezujících) podmínek a přístup může být i placený
- Zvážení všech etických, právních, licenčních a jiných omezení
- Princip "As open as possible, as closed as necessary" (Evropská komise)

Data Life Cycle: Reuse

- Použití dat pro jiný účel, než pro který byla nasbírána, např.:
  - Jako referenční data pro jiný výzkum
  - Ověření výsledků původního výzkumu
  - Propojování výsledků více studií do meta-studií

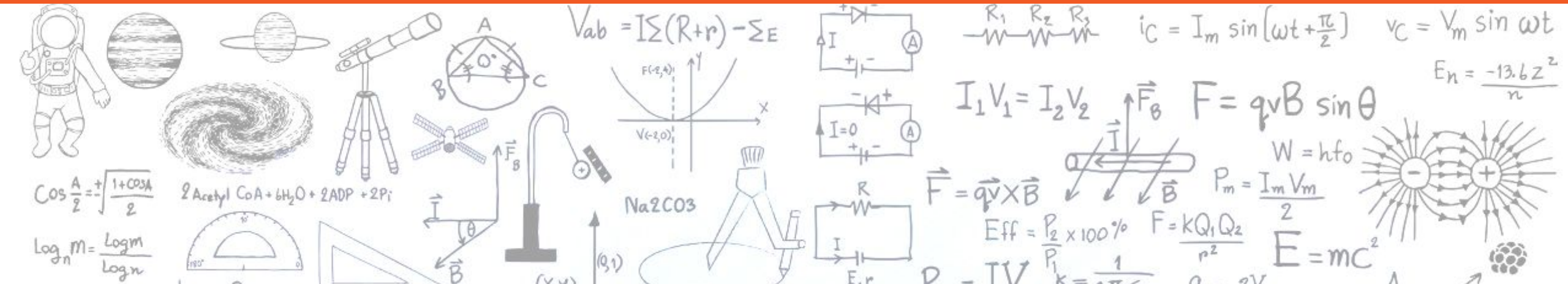
Role v RDM

- **Výzkumník (Researcher)** = pracuje s daty během výzkumu, píše DMP pro svůj projekt
  - Výzkumníci, PhD studenti, žadatelé o granty, projektoví manažeři, atd.
- **Správce dat (Data Steward)** = stará se o práci s daty na různých úrovních
  - **Policy** = spolupráce s vedením a grantovými agenturami; dohled na správný postup, etické a právní záležitosti
  - **Research** = spolupráce s vědci; kontrola a pomoc s tvorbou DMP
  - **Infrastructure** = spolupráce s IT; IT řešení pro RDM, infrastruktura, atd.
  - Někdy ještě jemnější dělení na různé role (data custodian, data curator, apod)

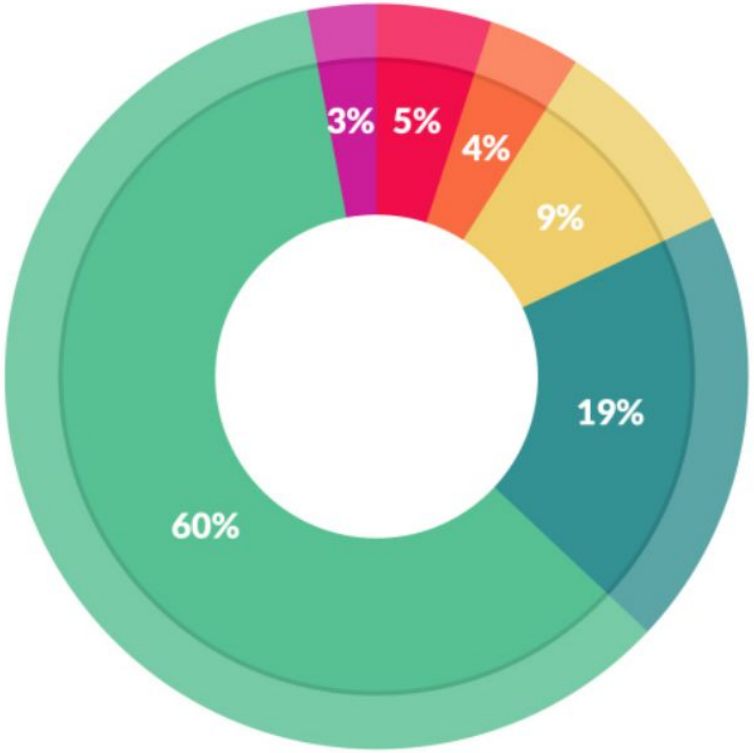
DSW

DATA STEWARDSHIP WIZARD

FAIR



Motivace



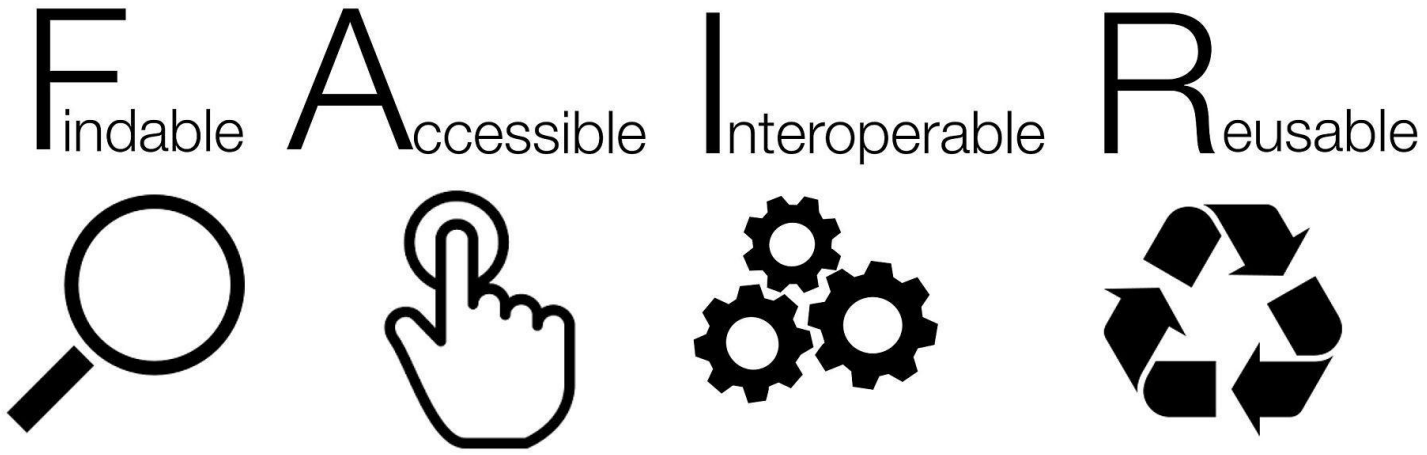
What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: Data Science Report 2016, CrowdFlower, 2016: http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

FAIR Principy

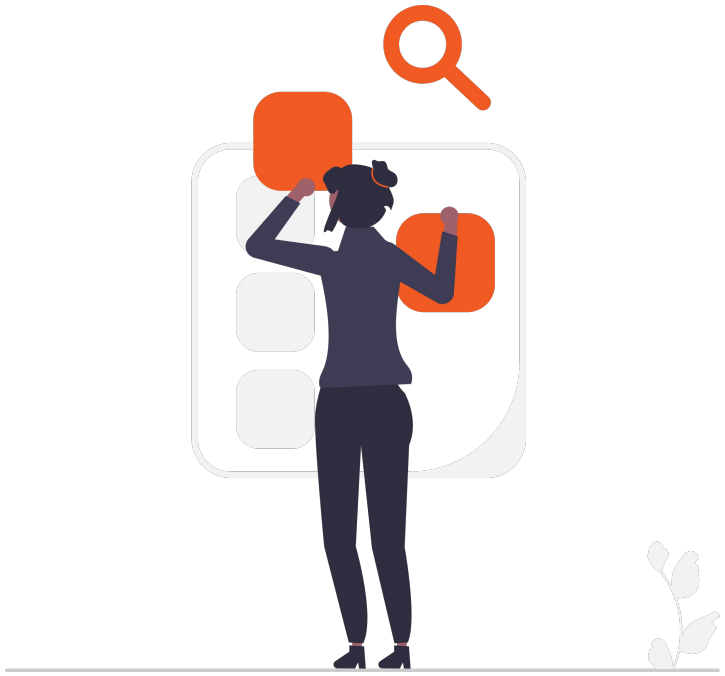
- Hlavním cílem je vylepšit **znovupoužitelnost vědeckých dat**
- Zdůrazňují strojovou zpracovatelnost (*machine-actionability*) a kvalitní metadata
- Formulováno 16 obecnými principy (r. 2016)



Findable (nalezitelná)

- Prvním krokem ke znovupoužitelnosti je nalezení dat

- (Meta)data mají globálně unikátní persistentní identifikátor
- Data jsou popsána bohatými metadaty
- (Meta)data jsou registrována v prohledávatelném zdroji



Accessible (dostupná)

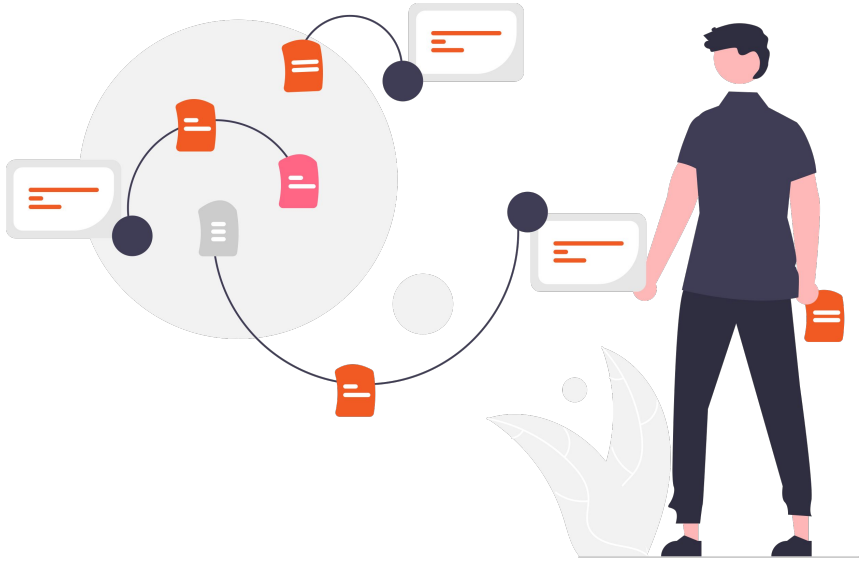
- Po nalezení dat musí být jasné, jakým způsobem je k nim možné přistupovat

- **Accessible ≠ Open**

- Přístup k (meta)datům pomocí standardního komunikačního protokolu
  - Použití autentizace a autorizace pokud je potřeba

Interoperable (interoperabilní)

- Možnost integrace dat s jinými daty
- Možnost procesování v různých aplikacích a workflow
- Použití standardních formátů, slovníků a ontologií (RDF, JSON-LD, OWL)



Reusable (znovupoužitelná)

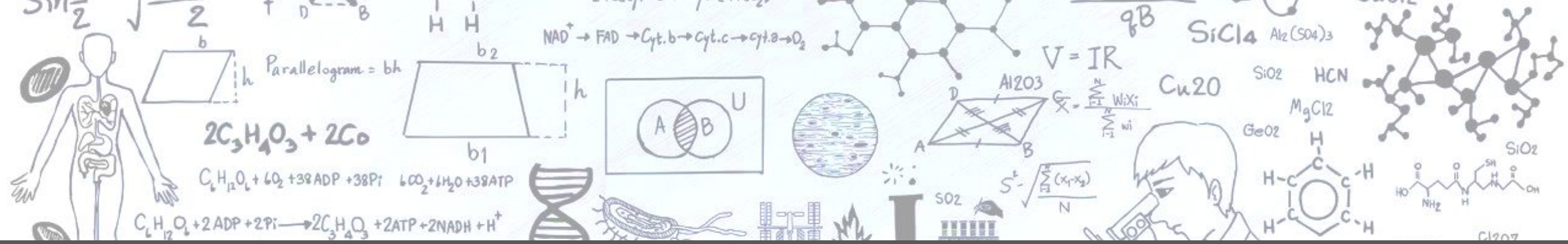
- Hlavním cílem FAIR je znovupoužitelnost
- Kvalitní popis (meta)dat
  - Licence
  - Původ *(provenance)*
  - Standardy komunity v dané doméně

Význam FAIR

- FAIR vize představuje "vrchol evoluce Homo Sapiens" v oblasti nakládání s daty
- Zaměřuje se na obecné technické a organizační aspekty
- Implementace FAIR principů je velké celosvětové téma se spoustou výzev – pionýrská doba se spoustou příležitostí

Kde se dozvědět více?

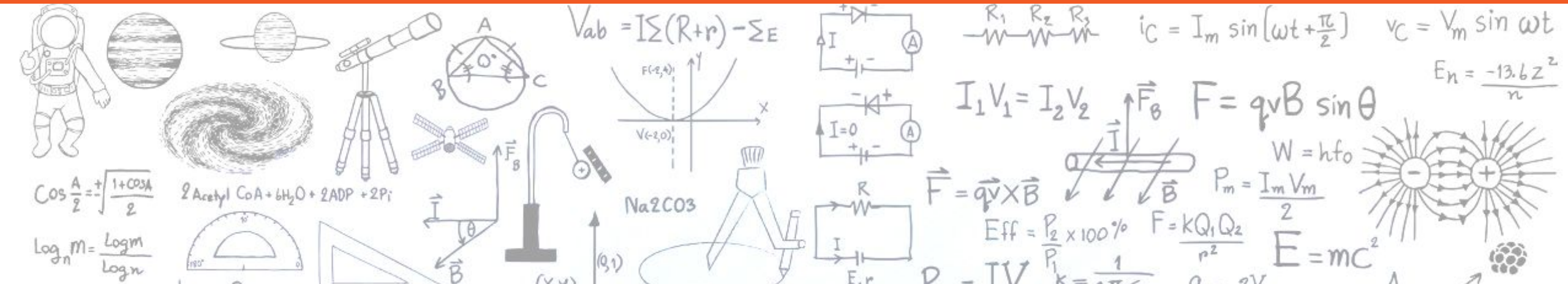
https://www.go-fair.org



DSW

DATA STEWARDSHIP WIZARD

Data Management Plan



DMP je běžná (a vyžadovaná) součást projektové žádosti

DSW

ZonMw

NIH National Institutes of Health

SCIENCE EUROPE

NSF

European Commission
Horizon Europe

erc
European Research Council

GAČR
GRANTOVÁ AGENTURA ČESKÉ REPUBLIKY

TA ČR

… a mnoho dalších

K čemu je dále dobré mít DMP

- Definuje role a zodpovědnosti pro práci s daty v týmu
- Pomůže při plánování prostředků a vybavení
- Pomůže identifikovat rizika a vybrat vhodná řešení na začátku projektu: "Prevence vs. hašení požárů"
- Usnadňuje sdílení, znovupoužitelnost a zachování dat (naplnění FAIR principů)
- I samotný proces plánování má svou hodnotu

  "Plány jsou ničím; plánování je vším."

Obsah DMP

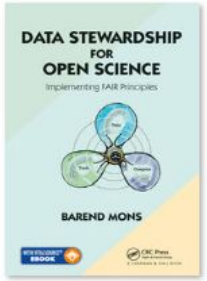
- Obecné informace o projektu
- Popis dat, která budou během projektu použita a vygenerována
- Použití metadat a ontologií, dokumentace
- Ukládání dat, bezpečnost a strategie pro zachování dat po ukončení projektu
- Sdílení dat
- Náklady a lidské zdroje potřebné pro správu dat
- Etické a právní otázky, licence
- Způsob naplnění FAIR principů

Jak vytvořit dobrý DMP?

Vytvořit dobrý DMP není snadné

Varianta 1: sám nastudovat vše potřebné

- Literatura
- Kurzy a tréninky

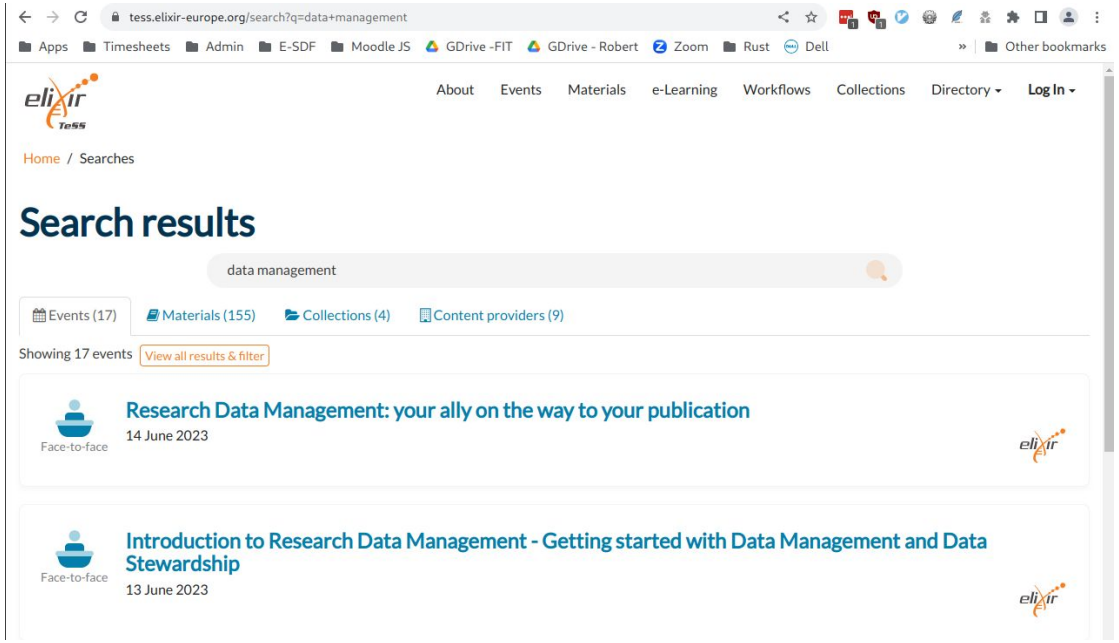


Book

Data Stewardship for Open Science

Implementing FAIR Principles
By Barend Mons

Edition	1st Edition
First Published	2018
eBook Published	27 February 2018
Pub. Location	New York
Imprint	Chapman and Hall/CRC
DOI	https://doi.org/10.1201/9781315380711
Pages	244
eBook ISBN	9781315380711
Subjects	Bioscience, Computer Science, Mathematics & Statistics

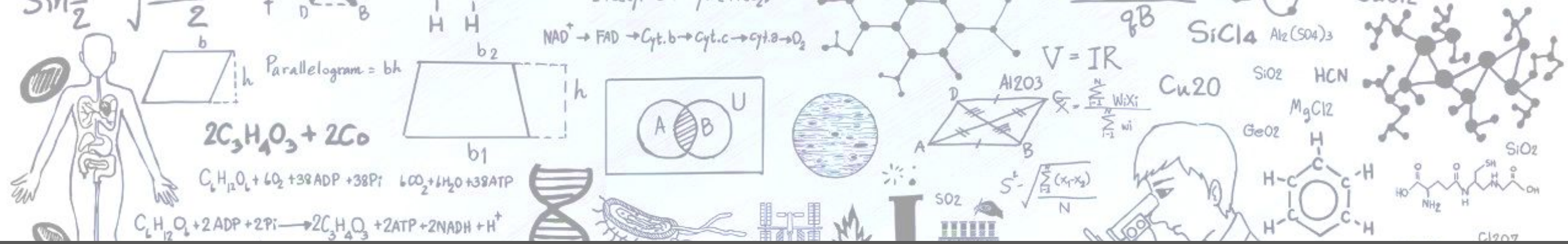


Jak vytvořit dobrý DMP?

Varianta 2: vyhledat pomoc

A. Institucionální data steward
   - Ne každá instituce zatím má
   - Bývá vytížen

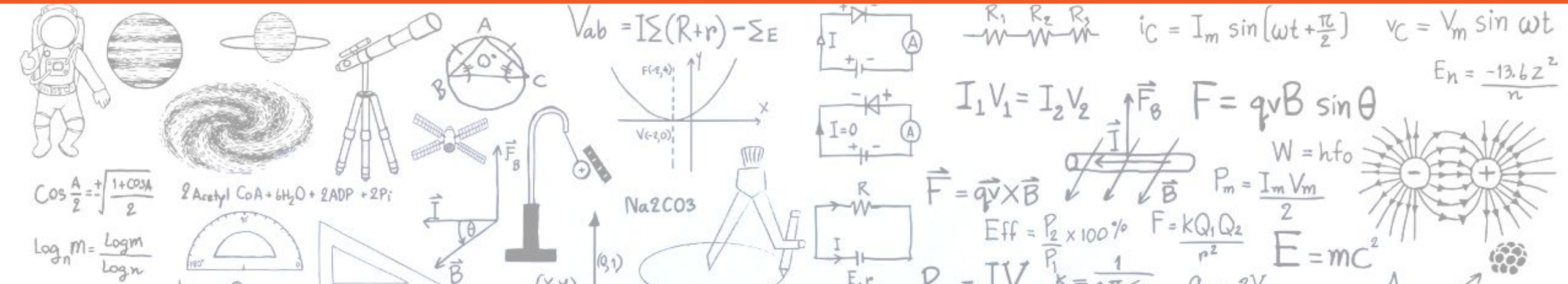
B. Data Stewardship Wizard = elektronický data steward se spoustou možností a funkcí, který je k dispozici 24x7



DSW

DATA STEWARDSHIP WIZARD

Data Stewardship Wizard



Data Stewardship Wizard

DSW

- Open-source vyvíjený ve spolupráci **ELIXIR CZ** a **ELIXIR NL**
- Expertní systém pro plánování správy dat a tvorbu data management plánů
  - *“From burden to benefit”*
- Doporučený nástroj mj. v **Horizon Europe Program Guide.**

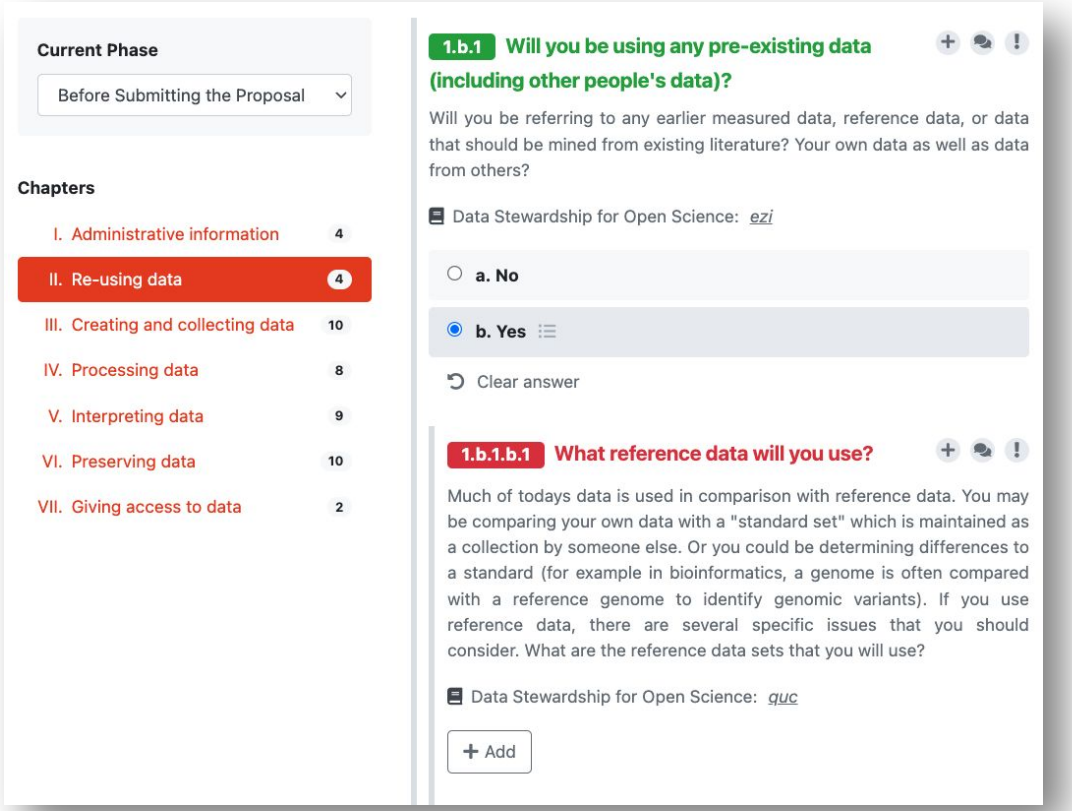
Hlavní myšlenky DSW

- Minimum psaní = plán není esej, psaní jen tam, kde je to nezbytné
- Vedení = DSW vede uživatele skrz tzv. *smart questionnaire*
- Flexibilita = lze upravovat obsah a integrovat s jinými službami
- Otevřenost = kdokoliv jej může využít a vytvářet vlastní obsah
- Orientace na uživatele = vývoj se výrazně řídí na základě zpětné vazby

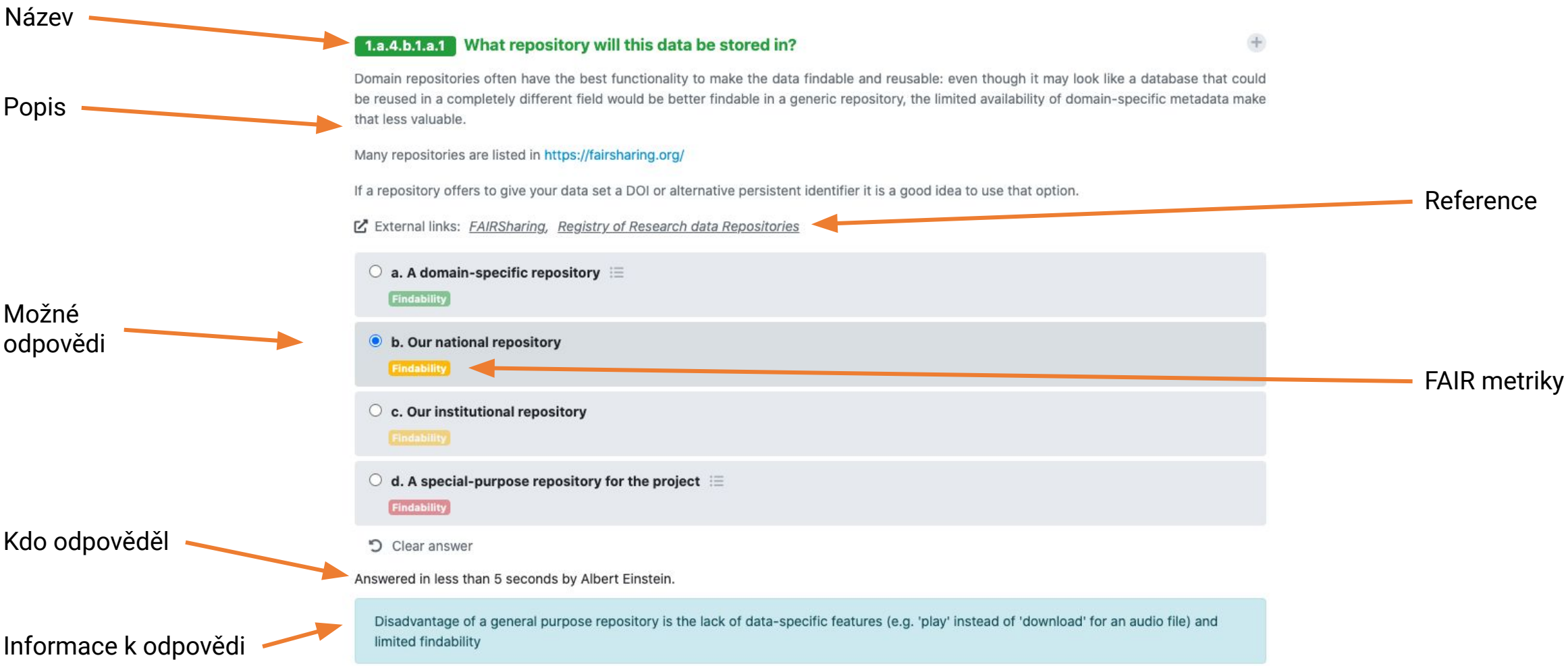
DMP jako předletová příprava

Dotazník (Smart Questionnaire)

- Interaktivní formulář pro získávání všech informací důležitých pro DMP
- Odpovídat je možné v libovolném pořadí
- Jsou zobrazovány pouze relevantní otázky na základě předchozích odpovědí



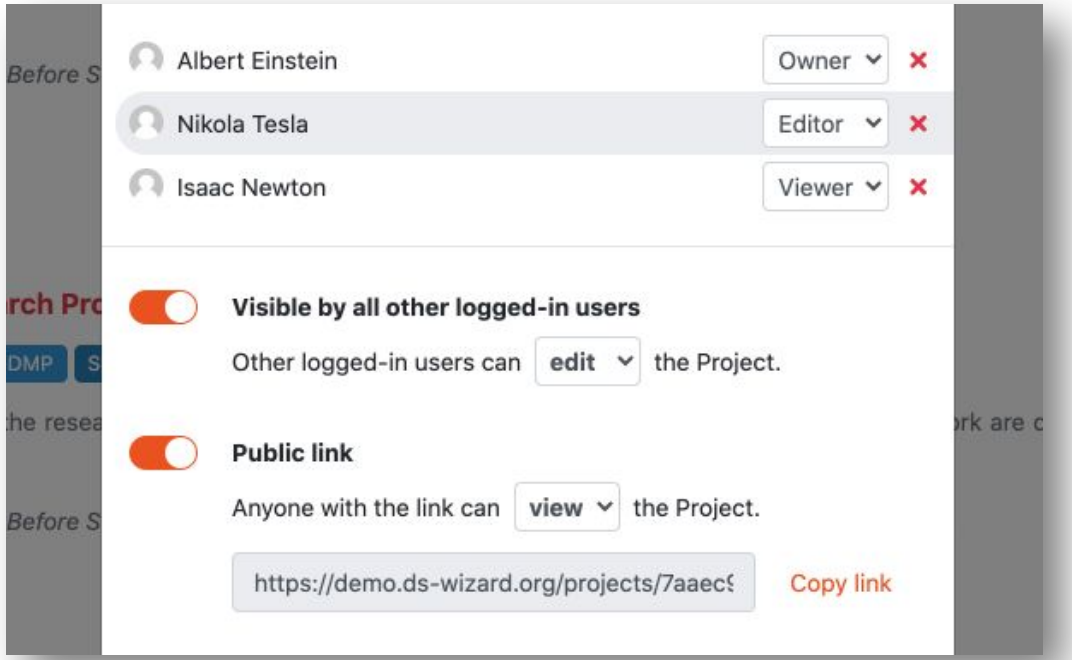
Otázka

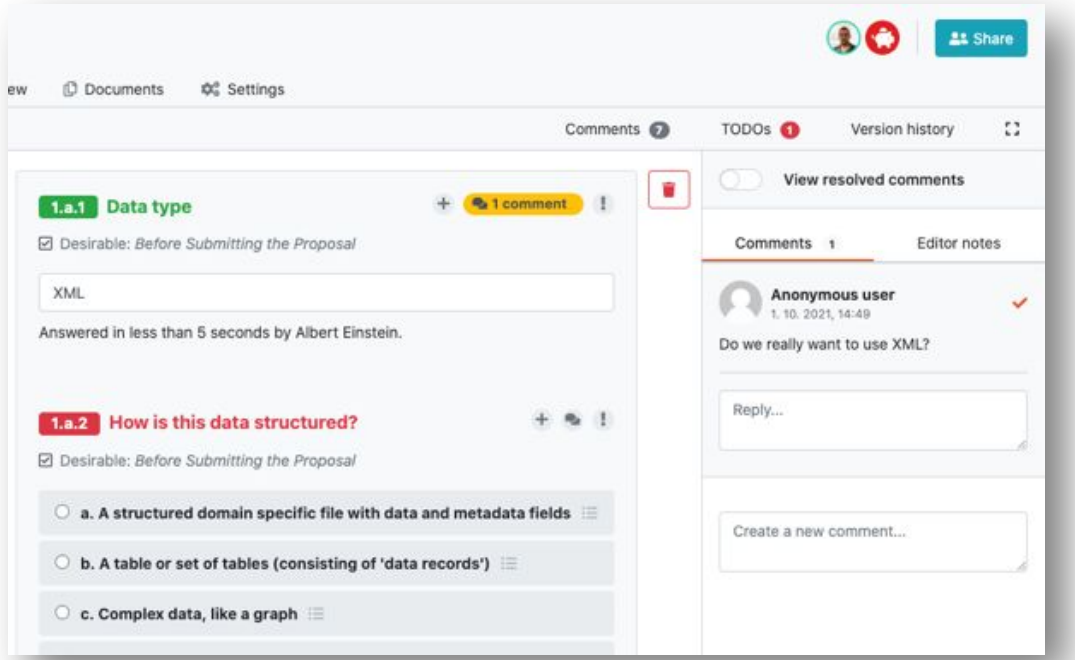

Název
Popis
Možné odpovědi
Kdo odpověděl
Informace k odpovědi
1.a.4.b.1.a.1 What repository will this data be stored in?
Domain repositories often have the best functionality to make the data findable and reusable: even though it may look like a database that could be reused in a completely different field would be better findable in a generic repository, the limited availability of domain-specific metadata make that less valuable.
Many repositories are listed in https://fairsharing.org/
If a repository offers to give your data set a DOI or alternative persistent identifier it is a good idea to use that option.
External links: FAIRSharing, Registry of Research data Repositories
Reference
a. A domain-specific repository
Findability
b. Our national repository
Findability
FAIR metriky
c. Our institutional repository
Findability
d. A special-purpose repository for the project
Findability
Clear answer
Answered in less than 5 seconds by Albert Einstein.
Disadvantage of a general purpose repository is the lack of data-specific features (e.g. 'play' instead of 'download' for an audio file) and limited findability

Metriky

- Nejsou vysloveně "dobré a špatné" odpovědi, některé volby ale mohou být vhodnější než jiné.
- DSW poskytuje zpětnou vazbu formou metrik - F, A, I, R, G, O
- Výsledná hodnota se počítá jako vážený průměr všech odpovědí, které danou metriku ovlivňují

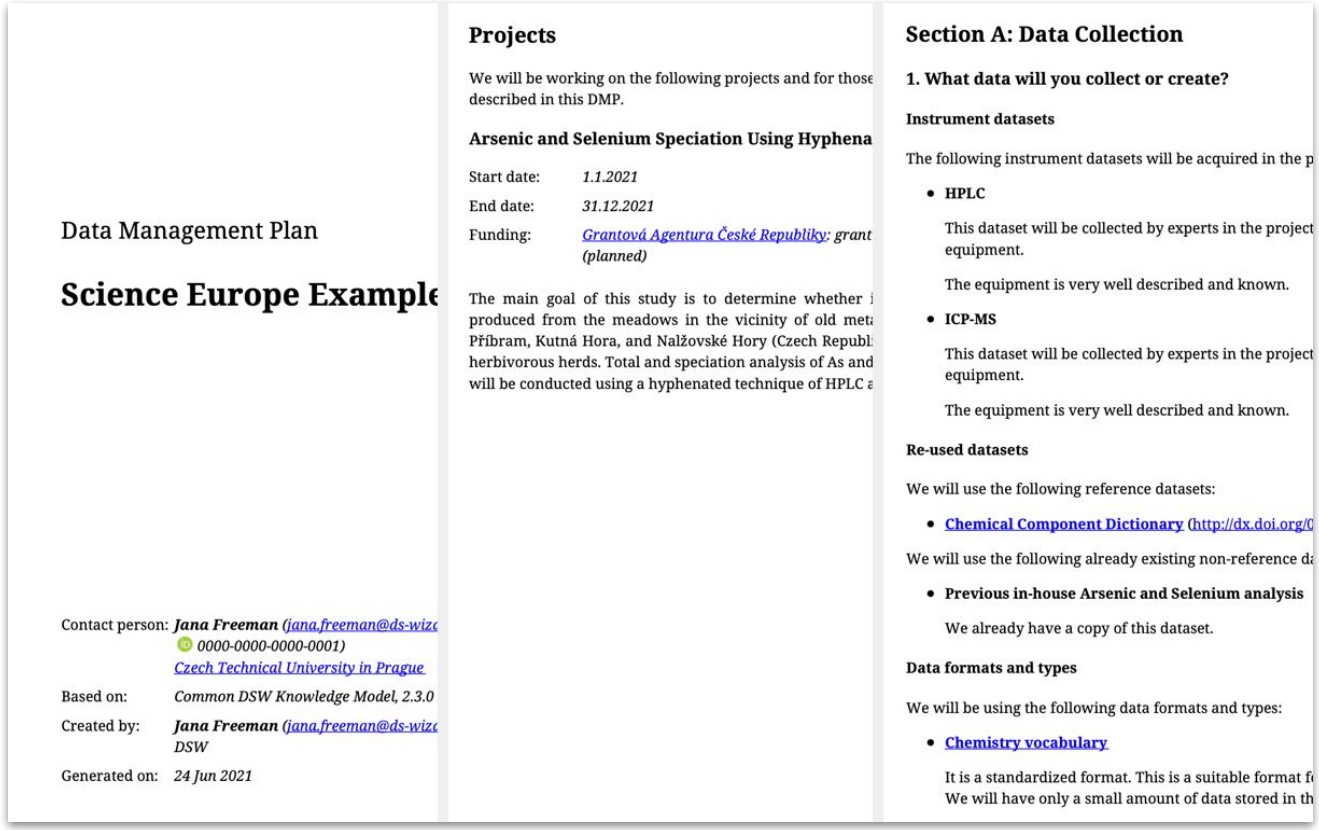
Online spolupráce





Vygenerování dokumentu s DMP

K dispozici je řada šablon na "push of button"



Data Management Plan

Science Europe Example

Contact person: ***Jana Freeman*** *(jana.freeman@ds-wiz*
0000-0000-0000-0001)
Czech Technical University in Prague
Based on: *Common DSW Knowledge Model, 2.3.0*
Created by: ***Jana Freeman*** *(jana.freeman@ds-wiz*
DSW
Generated on: *24 Jun 2021*

Projects

We will be working on the following projects and for those described in this DMP.

Arsenic and Selenium Speciation Using Hyphena

Start date: *1.1.2021*
End date: *31.12.2021*
Funding: *Grantová Agentura České Republiky: grant (planned)*

The main goal of this study is to determine whether i produced from the meadows in the vicinity of old met Příbram, Kutná Hora, and Nalžovské Hory (Czech Republ herbivorous herds. Total and speciation analysis of As and will be conducted using a hyphenated technique of HPLC a

Section A: Data Collection

1. What data will you collect or create?

Instrument datasets

The following instrument datasets will be acquired in the p

- **HPLC**

  This dataset will be collected by experts in the project equipment.

  The equipment is very well described and known.

- **ICP-MS**

  This dataset will be collected by experts in the project equipment.

  The equipment is very well described and known.

Re-used datasets

We will use the following reference datasets:

- **Chemical Component Dictionary** (http://dx.doi.org/0

We will use the following already existing non-reference da

- **Previous in-house Arsenic and Selenium analysis**

  We already have a copy of this dataset.

Data formats and types

We will be using the following data formats and types:

- **Chemistry vocabulary**

  It is a standardized format. This is a suitable format f
  We will have only a small amount of data stored in th

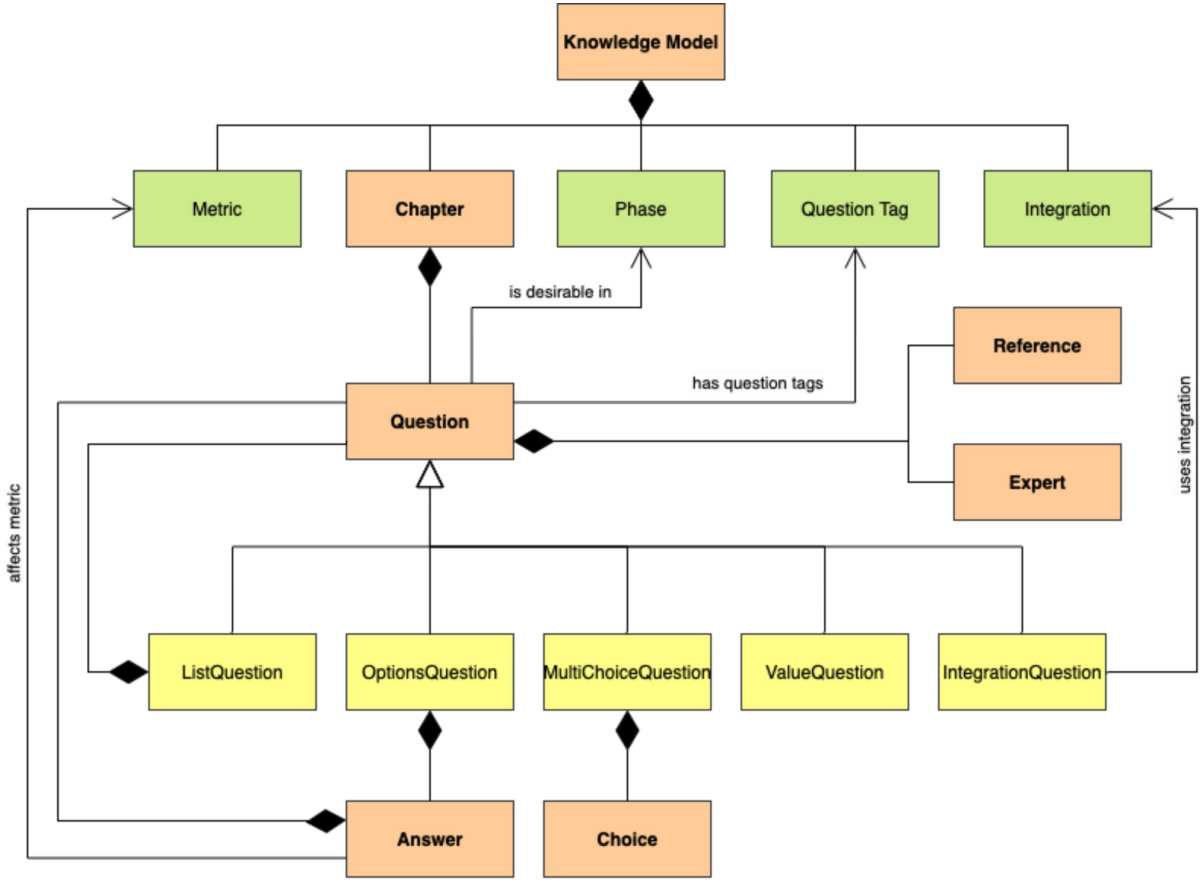
- Horizon 2020 DMP
- Horizon Europe DMP
- Science Europe DMP
- Machine-actionable DMP (RDA Common Standard)
- Questionnaire Report

Jak to funguje

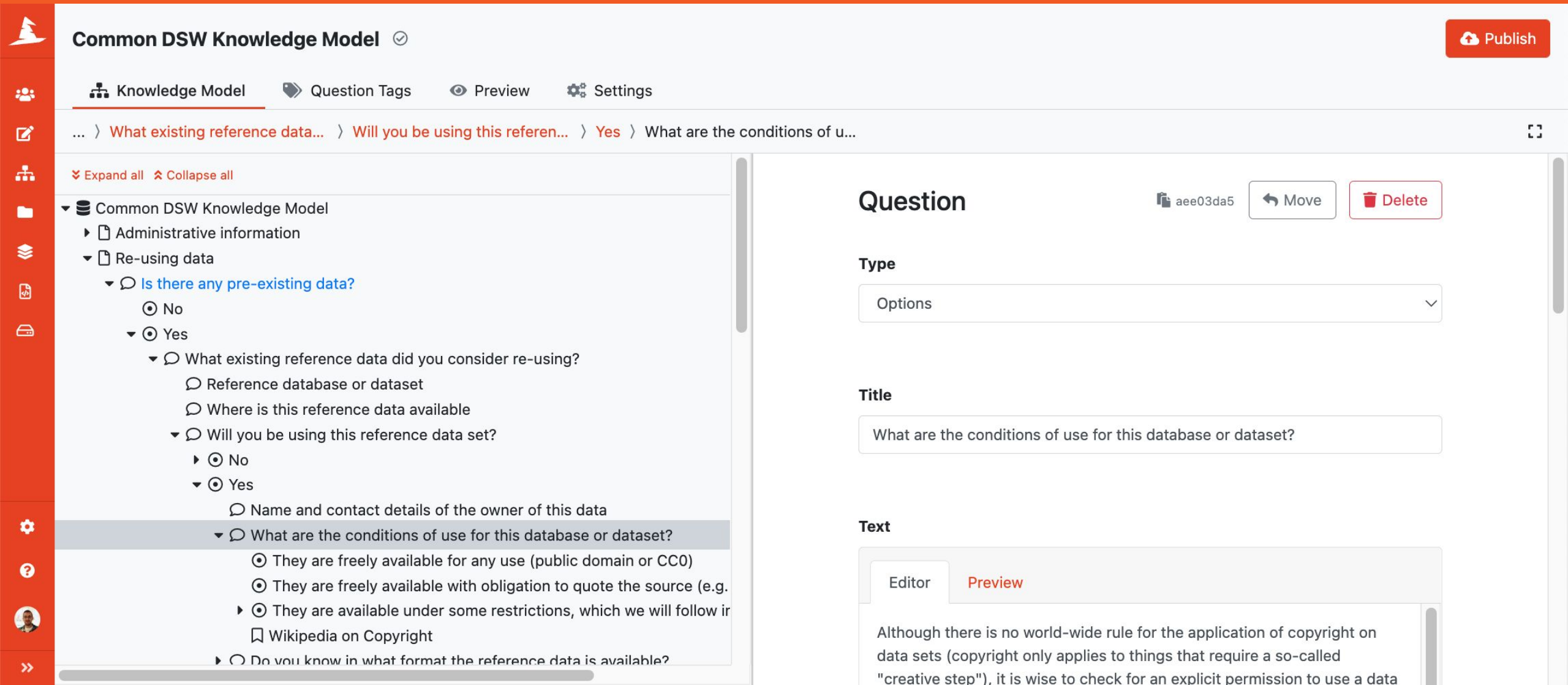
Knowledge Model

Knowledge Model (KM)

- Obsahuje **znalosti** o tom, na co a jak je potřeba se ptát
- **Šablona** pro strukturovaný dotazník (questionnaire)
- Stromová struktura sestávající z **kapitol**, **otázek**, **odpovědí**, navazujících otázek a dalších zdrojů



Knowledge Model Editor



Typy otázek

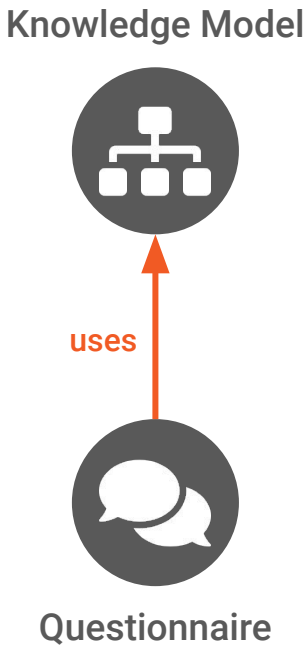
- Hodnota (Value) – text, email, datum, číslo,...
- Výběr odpovědi (Options) – umožňuje větvení dotazníku
- Výběr z více možností (Multi-Choice)
- Seznam položek (List of items) – každá položka má stejné odpovědi
- Integrace (Integration) – propojení s externí službou

Možnosti přizpůsobení KM

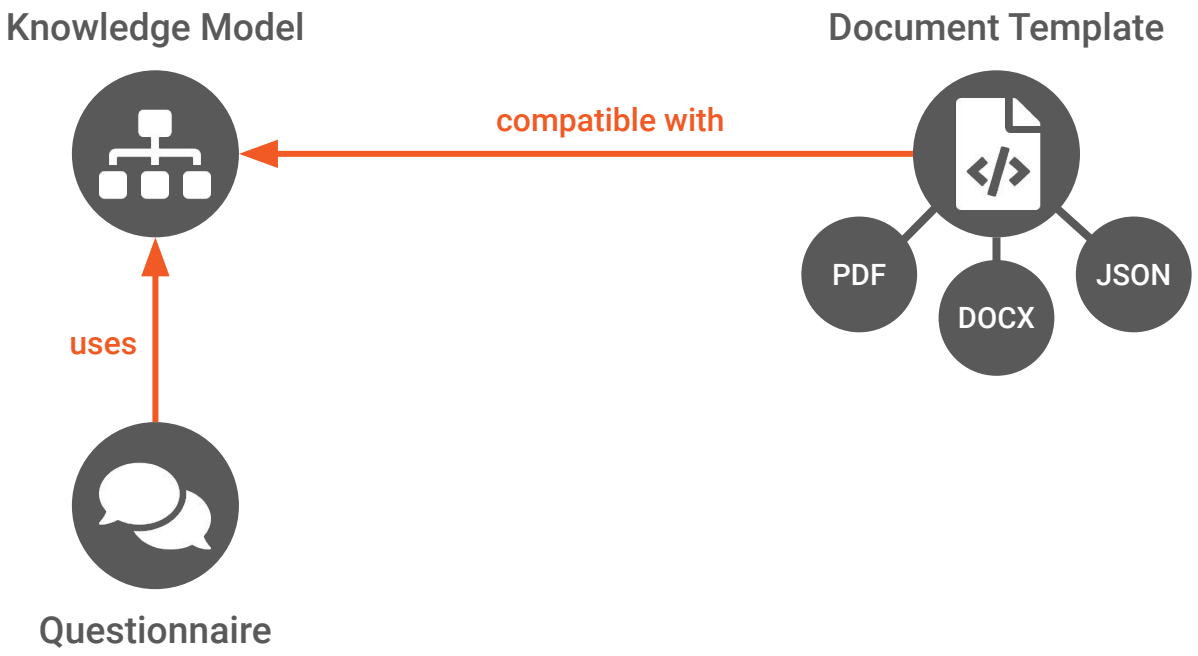
- Rozšíření nebo úpravy existujících KM
- Vhodné pro rozšíření pro jiné domény nebo instituce

- Příklad
  - **Common DSW Knowledge Model** – obecný KM pro data stewardship
  - **Life Sciences DSW Knowledge Model** – rozšiřuje obecný KM o otázky z Life Sciences domény

Data Stewardship Wizard

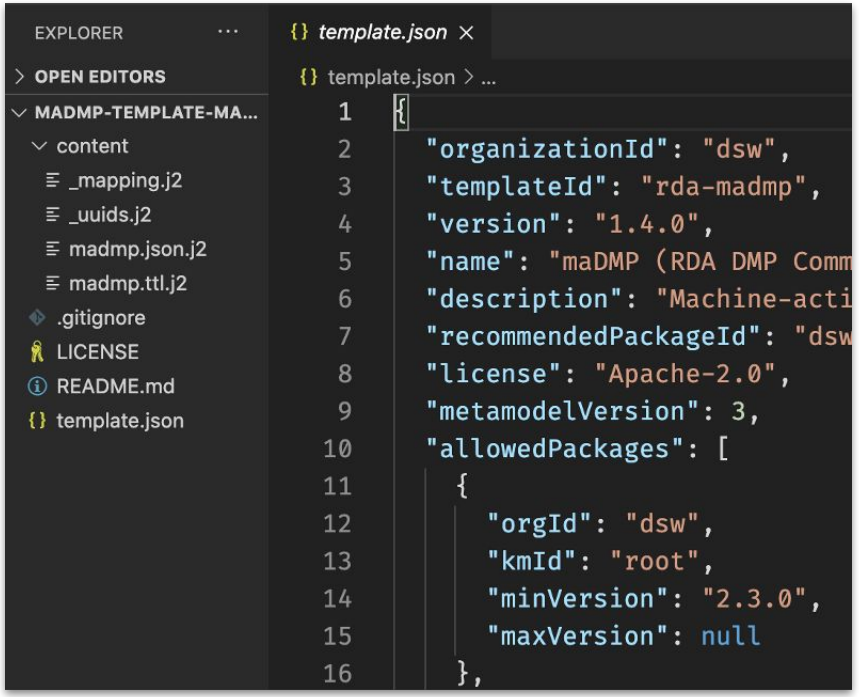


Data Stewardship Wizard

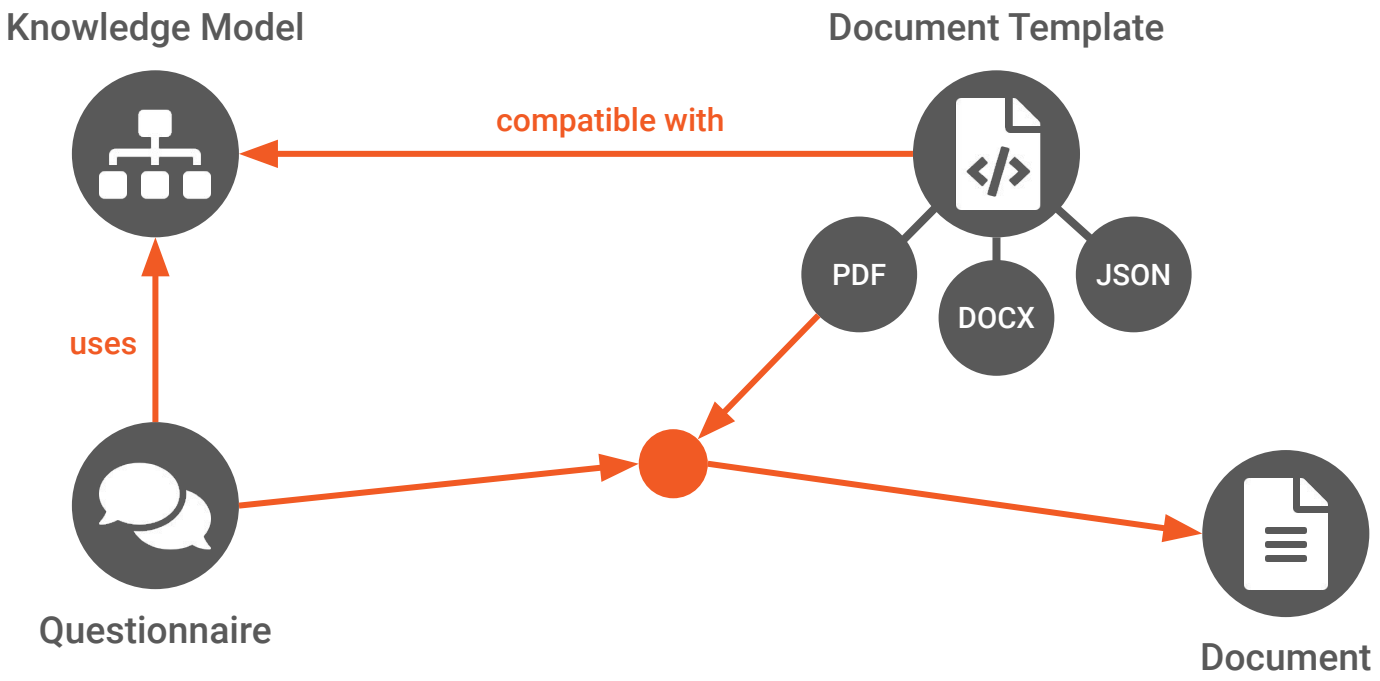


Šablona pro dokumenty (Document Template)

- Tvoření šablon vyžaduje více **technických dovedností**
- Šablony jsou složené z **JSON** metadat, **Jinja2** šablon a dalších souborů
- Tvoření usnadňuje **DSW Tempalte Development Kit** (TDK)

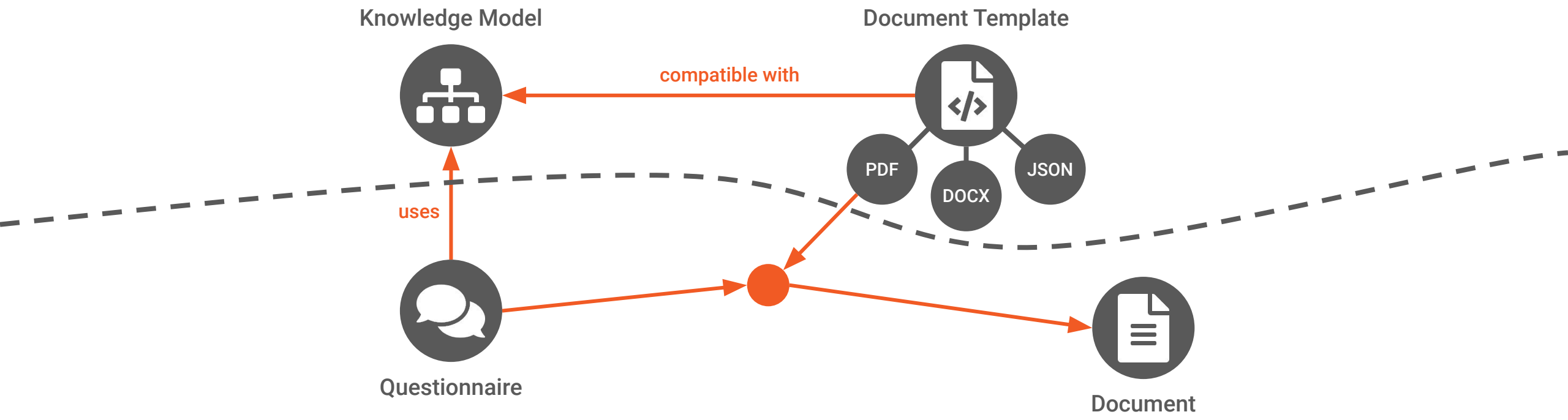


Data Stewardship Wizard



Data Stewardship Wizard

Data Steward



Researcher

A mnohem víc

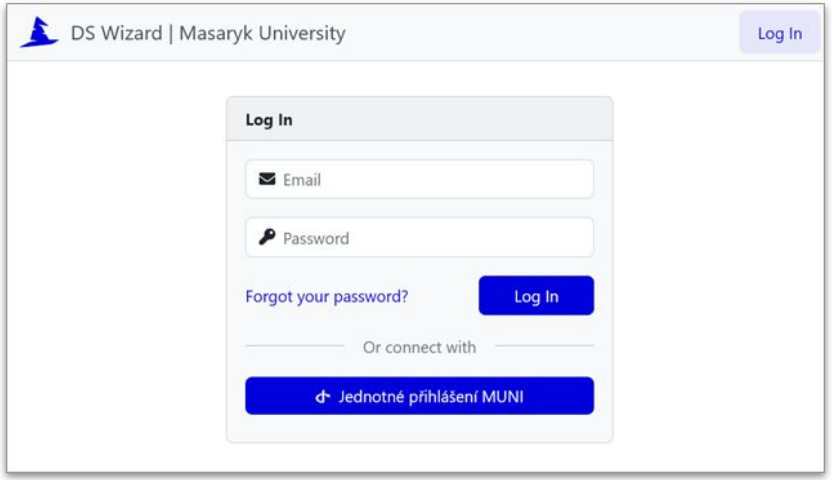
- Data stewardship hints
- Project phases
- Project templates
- Editor notes
- Version history
- Migrace knowledge modelů & aktualizace dotazníků
- Integration widget
- Submission service
- Single-Sign On
- ...

Kde se DSW používá

DSW

DSW a další na Masarykově univerzitě

- DSW MUNI – univerzitní instance DSW: https://dsw.muni.cz/
- Open Science web MU: https://openscience.muni.cz/
  - Návod na použití DSW MU s příklady DMP z různých oborů.
  - Kontakty na Open Science core tým a fakultní metodiky: https://openscience.muni.cz/kontakty
- Vzdělávací aktivity:
  - Kurz Bezpečná správa výzkumných dat (chystáme třetí běh kurzu; datum bude upřesněno).
    - Primárně pro zájemce z řad výzkumných pracovníků, projektové a vědecké podpory.
  - Kurz CORE042: Data – odpověď na základní otázku života, vesmíru a vůbec…
    - Primárně určen pro bakalářské studenty všech fakult MU.
    - Vytvořeno jako zcela nový předmět v rámci vznikajících kurzů Společného univerzitního základu pro rozšiřování znalostí studentů za hranice jejich domovského oboru.
      - Chceme zasévat semínka pro výchovu nové generace Data Stewards.



Poděkování

Provoz a vývoj nástroje DSW je podporován infrastrukturou ELIXIR CZ (MŠMT grant č.: LM2023055).

Děkuji za pozornost

- **Otázky a diskuse**

- Data Stewardship Wizard

  https://ds-wizard.org

  @dswizard_org

- Robert Pergl

  robert.pergl@ds-wizard.org